Click here to view PowerPoint presentation; Press Esc to exit

Cataloging and Indexing of Electronic Information Resources

Gail Hodge
Information International Associates, Inc.
312 Walnut Place
Havertown, Pennsylvania 19083, USA

Abstract

The rationale for cataloging and indexing of electronic information is much the same as for print materials. Cataloging and indexing provide a surrogate for the item, which facilitates resource discovery and access. But, what has changed in the electronic information environment is the terminology. In the Internet environment, the terms cataloging and indexing have been replaced with the term "metadata." Metadata is often defined as "data about data" or "information about information." The term, which originated with the data and computer science communities, is now in general use for the cataloging and indexing of electronic information sources.

Metadata serves three general purposes. It supports resource discovery and locates the actual digital resource by inclusion of a digital identifier. As the number of electronic resources grows, metadata is used to create aggregate sites, bringing similar resources together and distinguishing dissimilar resources.

There are a variety of metadata schemes that serve different purposes for different object types, subjects and audiences, including the Dublin Core, the Global Information Locator Service, the Text Encoding Initiative, Header,the Encoded Archival Description, the Online Information Exchange International, the Content Standard for Digital Geospatial Metadata, the Data Documentation Initiative, and the draft Technical Standard for Still Images. A metadata scheme has three components -- semantics, content and syntax. An extension adds elements to an existing scheme to describe a particular resource type, handle material on a particular subject, or address the needs of a particular user community. Profiles are subsets of a larger scheme that are implemented by a particular user community. Metadata can be embedded in an electronic resource or stored in a separate file.

A growing number of tools are available to create and edit metadata. Creation may be done manually or by metadata generators that extract key information from the object. While many projects aimed at having metadata created by the object's author, this has proved to be difficult to implement. An alternative is to have a core set of metadata created by the author with editing and quality control performed by a librarian or editor that has a view to the whole collection.

With disparate metadata schemes, ensuring that information collected in a specific scheme by one organization for a particular purpose can be exchanged, transferred or used by another organization for a different purpose becomes an issue. Metadata frameworks, crosswalks, and registries are ways to achieve interoperability.

Use of controlled and uncontrolled vocabulary terms is encouraged, particularly within specific subject domains. However, most metadata schemes do not dictate the use of a particular controlled vocabulary but instead allow the vocabulary scheme to be defined within the syntax.

In order to increase the use of metadata, systems that support metadata creation and search engines that take better advantage of metadata must be developed. Communities of practice should develop content standards, along with other groups that share common interests. Stakeholder groups must be made aware of the importance of metadata for the short and long-term enhancement of the electronic environment.

*Paper presented at the RTO IMC Lecture Series on "Electronic Information Management for PfP Nations",
held in Vilnius, Lithuania, 24-26 September 2002, and published in RTO-EN-026.*

1.0 The purpose of metadata

Similar to traditional cataloging and indexing, metadata performs three main functions. It facilitates discovery of relevant information, locates the specific resource, and organizes electronic resources into collections.

1.1 Resource discovery

One of metadata's primary functions is to support resource discovery by describing aspects of the original electronic resource in which the designated user community may be interested. Metadata, such as titles, subject terms and abstracts or descriptions, are particularly important for electronic resources, such as datasets or photographs, that have little if any text content on which current text-based Web searching can be performed.

Metadata can describe the resource at any level of aggregation -- a single resource; a part of a larger resource, for example, a photograph in an article; or a collection of resources, such as a digital library. The level at which metadata is applied depends on the type of data and the anticipated access needs. Datasets are generally cataloged at the file or collection level. Electronic journal articles may be cataloged individually, sometimes with no concern for metadata at the issue or journal title levels. Generally, the metadata for Web sites is applied to one or more pages that make up a cohesive resource with informational value.

1.2 Location of electronic resources

Metadata supports the location of the actual digital resource on the network. Most metadata schemes include an element that is defined as the unique identifier needed to locate the resource.

In practice, most metadata schemes continue to use the URL, or the Uniform Resource Locator, as the unique identifier. The URL is the physical address, server, directory and file name for the resource. This provides fast look-up but is problematic as the Web grows and information managers need to move the physical locations of the resources. In the case of electronic journals, URL changes may occur due to the merger or acquisition of one publisher by another. URLs that are not up-to-date result in the famous 404 message indicating that the Web page cannot be found.

In an effort to solve this problem, two major systems have been developed. First, OCLC developed the Persistent URL. This method continues to use the URL construct, but it sets up a resolver service. The PURL is used in the metadata record or in reference links that refer to the electronic resource. When a browser attempts to locate the PURL, it accesses the record in the PURL Resolver service at OCLC. The Resolver uses standard Internet redirection to access the actual URL of the resource's physical file location. If the location for the actual page changes, its owner must change the URL in the Resolver, but the PURL that has been published remains the same.

The PURL is structured as:

http://purl.oclc.org/[specific resource file name]

The beginning of the PURL is the URL for the PURL Resolver Service (in the example above, the resolver at OCLC is used) and the file name in brackets is the file name for the specific resource.

The second method is the Handle System® developed by the Corporation for National Research Initiatives (CNRI) under contract to several U.S. government agencies. In the Handle, the prefix is a unique identifier assigned to the resource owner by the central Handle System. This prefix ensures that the identifier is unique. Following the slash is the suffix assigned to the item by the producer.

A Handle is structured as:

[unique prefix for the assigning agency]/[persistent, unique identifier for the resource]

The unique identifier in a Handle can be any item ID. Possibilities include the ISBN, the Standard Item Contribution Identifier (SICI), the Publisher Item Identifier (PII), or a local accession number.

The Handle also uses a resolver service, but it allows more flexibility in the structure and syntax of the identifier. Because it actually uses a database scheme, a single Handle can resolve to multiple locations for different versions of the same resource. Different versions of an electronic resource, for example one in HTML and the other in pdf, can be uniquely identified even though they have the same Handle, because the database also contains the data type. The data types can be resolved based on a user's preference or an interface can be designed that offers the user a choice between the versions.

The Handle is the underlying technology for the Digital Object Identifier. The DOI, managed by the International DOI Foundation, establishes a specific syntax for the DOI under the Handle framework. The DOI is the basis for a system called CrossRef. CrossRef is a DOI Registration Agency formed by a consortium of electronic journal publishers. The members of CrossRef deposit their DOIs into a central repository maintained by CrossRef. The purpose of CrossRef is to facilitate linking between electronic journals, primarily from the references at the end of an article to the full text for those articles. The DOI in CrossRef is used to form the reference link from a reference to the full text article of another publisher.

As mentioned earlier, CrossRef is a DOI Registration Agency, which maintains a central repository of DOIs in order to allow publishers to move their physical files, while maintaining a persistent link in previously published references. In addition to the DOI itself, CrossRef maintains a minimal set of metadata for each DOI. This limited metadata, consisting of the article title, the first author's last name, and journal citation information, allows a publisher or library to find the DOI for an article published by a member of the CrossRef system in order to embed the DOI in a reference.

1.3 Organization of resources into collections

In addition to the discovery of specific resources, metadata brings similar resources together and distinguishes dissimilar resources. As the number of Web-based resources grows exponentially, aggregate sites, portals, or subject gateways are increasingly useful in organizing links to resources based on audience or topic. Such lists can be built as static Web pages, with the names and locations of the resources "hard coded" in the HTML. However, it is more efficient and increasingly more common to build these pages dynamically from metadata stored in databases.

Another method of organizing Web information is through channels. Channels are pre-selected Web sites that automatically "push" collections of information to a user's browser. They are commonly used for continuously updated information such as stock quotes and news. The dominant metadata scheme for webcasting is the Channel Definition Format (CDF) developed by Microsoft and its partners. The CDF provides metadata elements such as the title of the channel, an abstract, the publication date, the last date the content was modified, the logo for the channel and the schedule on which the channel's content is updated so the "pushing" can be scheduled.

2.0 Basic metadata structure

This section describes the general structure of a metadata scheme, the modification of a scheme to increase its flexibility and usefulness by various communities of practice, and the storage of metadata.

2.1 Components of a metadata scheme

A metadata scheme (also called schema) is made up of three structural components – semantics, content and syntax. The definition or meaning of the elements is known as the semantics, and include the tag set for the elements. For example, a scheme for a text resource may define the Title element with a tag of TI. Generally, the semantics of a metadata scheme are grouped into three types -- descriptive, structural, and administrative. Descriptive metadata identifies a resource for purposes of discovery and identification. It includes elements such as title, abstract, author, and keywords. Administrative metadata provides information to help manage a resource, such as when and how it was created, its file type and other technical information. Rights management metadata is a subset of administrative metadata dealing with intellectual property rights and who can access the resource. Structural metadata indicates how compound objects are put together or how this resource relates to others in the collection.

The set of preservation metadata currently being developed by the Research Libraries Group and OCLC includes elements from all three of these semantic types, but it adds elements specific to preservation activities such as the provenance of the item, the preservation strategies employed, its migration history, etc. (Preservation metadata is discussed in more detail in the session on "Archiving and Preservation in Electronic Libraries.")

The scheme may also specify syntax rules for how the elements and their content should be encoded. Metadata can be encoded in MARC21, in "keyword=value" pairs, or in any other definable syntax. Many current metadata schemes use XML (Extensible Mark-up Language). A metadata scheme with no prescribed encoding syntax is called "syntax independent."

The third structural component of a metadata scheme is the content, or the values used to complete the elements. A scheme may specify rules, also called a "content standard," for the formulation of the content (for example, how to identify the title) or rules for the representation of the content (for example, capitalization, language or transliteration rules).

2.2 Extensions and profiles

Specific implementations or the needs of a certain community can result in modifications to a metadata scheme. Since it is often difficult to anticipate the ways in which a scheme might be used, schemes that can easily be modified are preferred over those that are more restrictive. Modifications are of two types: extensions and profiles.

An extension is the addition of elements to an already developed scheme to support the description of a particular resource type, to handle material on a particular subject, or to address the needs of a particular user community. Profiles are subsets of a larger scheme that are implemented by a particular user community. Extensions generally increase the number of elements that can be used; profiles constrain the number of elements, refine or narrow the definitions of certain elements, or specify the rules for completing the content of certain elements.

In practice, many applications use both extensions and profiles of base metadata schemes. The metadata scheme for the U.S. Department of Education's Gateway to Educational Materials (GEM) Project is based on the Dublin Core. However, GEM limits the elements to be used (for example, Contributor is not used). It also extends the Dublin Core element set by adding elements that are important to the educational community when describing and using educational resources. These fields include audience (teacher versus student), grade level, and relevant educational standards.

Similarly, the Visual Resources Association (VRA) has established core categories (or elements) to describe visual materials such as buildings, photographs, paintings and sculptures in visual resource collections of slides or photographs. Therefore, metadata for these materials must accommodate the description of the same resource in different media, for example, the original painting, a slide of the painting, and a digitized image of the slide. The VRA Core Category scheme, a profile and extension of the Dublin Core, consists of 17 optional metadata elements: record type, type, title, measurements, material, technique, creator, date, location, ID number, style/period, culture, subject, relation, description, source, and rights. The Dublin Core Relation field is used to relate the records for the same resource in different media. The VRA Core scheme does not specify any particular syntax or rules for representing content. Managers of visual resource collections hope that use of the VRA Core Categories will allow them to share descriptions of original works as well as to better describe materials in their own collections.

2.3 Metadata Storage

Metadata can be embedded in an electronic resource or stored separately. For example, metadata is often embedded in HTML documents as metatags or in the headers of image files. The use of HTML metatags specifically makes the content of the metadata accessible to a variety of Web search engines. Storing metadata with the resource ensures the metadata will not be lost, eliminates problems of broken links between the resource and its metadata, and facilitates updating of the metadata and the resource.

However, sometimes it is difficult to embed metadata in certain types of resources. In these cases, storing metadata separate from its electronic resource simplifies the management of the metadata and facilitates search and retrieval. Metadata is often stored in a Web-accessible database system (for example a clearinghouse or library catalog) and then linked to the electronic information it describes by a URL or other identifier in the metadata.

3.0 Metadata schemes

Metadata schemes (also called "schema") have been developed and defined by a variety of communities, for different purposes, and for different types of electronic resources. This section describes some common metadata schemes. In addition, some lesser known schemes have been selected to show the range of electronic resources and purposes for which schemes have been developed. While the focus here is on electronic information resources, it should be noted that many other metadata schemes have been developed in support of e-commerce and electronic data exchange.

3.1 Dublin Core

The Dublin Core is perhaps the most well known metadata element set. The original objective of the Dublin Core was to define a set of elements that could be used by authors to describe their own Web resources. A few relevant elements and simple rules were defined so that non-catalogers could provide basic information for resource discovery.

The final Dublin Core consists of 15 elements: title, subject, description, source, language, relation, coverage, creator, publisher, contributor, rights, date, type, format, and identifier. All Dublin Core elements are optional and all are repeatable. The elements may be presented in any order. Note that in the following example relation, contributor and source are not applicable and so do not appear.

Dublin Core Elements For This Paper

Title: Cataloging and Indexing of Electronic Information Resources

Creator: Hodge, Gail

Subject: metadata

Description: Describes metadata standards and projects.

Publisher: NATO

Date: 20020601

Type: Text.Report

Format: text/html

Identifier: http://www...............

Language: en

Coverage.Spatial: International

Rights: Copyright 2002, Gail Hodge

While the Dublin Core description recommends the use of controlled values for fields where they are appropriate (for example, controlled terms from a thesaurus for the Subject field), this is not required. The content rules are up to the particular implementation, but the adoption of profiles that define domain-specific rules is encouraged.

The Dublin Core was developed to provide simple and concise descriptions specifically to support the resource discovery of Web-based documents. However, in part because of its simplicity, the Dublin Core has been used with other types of materials and for applications demanding increased complexity. The desire to be able to specify more detail resulted in unqualified (or simple) Dublin Core versus qualified Dublin Core. In qualified Dublin Core, qualifiers are used to refine the meaning of an element or to specify the domain values or rules for representing an element. The element "Date", for example, can be used with the qualifier "created" to narrow the meaning of the element to the date the resource was created. A qualifier can also be used in the element "Date" to specify the ISO 8601 standard as the required format for representing date.

There are hundreds or perhaps thousands of projects worldwide that use the Dublin Core for cataloging or to collect data from the Internet; more than fifty of these projects are linked from the Dublin Core Web site. The subjects range from cultural heritage and art to math and physics. Dublin Core is the basis for the Cooperative Online Resource Catalog (CORC) system at OCLC by which libraries can share catalog records for Web-based resources.

3.2 Global Information Locator Service (GILS)

GILS was developed by the U.S. government as a tool for enhancing public access to government information. Originally called the "Government Information Locator Service", GILS in various forms has been adopted by other governments and for international projects, leading to its current name, "Global Information Locator Service". International implementers of GILS include Australia, Germany, Singapore, and Hong Kong. GILS is also widely used with spatial and environmental clearinghouses implemented by countries and international organizations.

GILS specifies a profile of the Z39.50 protocol for distributed search and retrieval which is a common standard used in online library catalogs. It specifies the attributes (or the elements) that must be able to be searched in order for a system to be GILS compliant. However, organizations have specifically defined GILS elements for their own communities.

Since the purpose of GILS is to act as a locator service, GILS elements emphasize availability and distribution rather than description. Therefore, a GILS record may have elements such as the name and address of the distributor and information on ordering process.

A U.S. Federal GILS Core Record For This Paper

Title: Cataloging and Indexing of Electronic Information Resources

Originator: Gail Hodge

Local Subject Term: Metadata

Abstract: Describes metadata standards and projects for librarians.

Purpose: To serve as an educational aid to librarians, information center managers and others involved in the dissemination and creation of electronic resources.

Availability:

Distributor :

Name: Information International Associates (IIa)

Street Address: 122 S. Jefferson Circle

City: Oak Ridge

State: TN

Country: USA

Zip Code: 37830

Telephone: 865-481-0388

Fax: 865-481-0390

Order Process: This paper is available without charge by writing to IIa at the address provided.

The original goal of GILS was to provide high-level locator records for government resources, both electronic and non-electronic. GILS records were intended to describe aggregates or collections such as catalogs, publishing services and databases. However, some organizations use GILS at the individual item (journal article or technical report) level.

3.3 Text Encoding Initiative (TEI) Header

The Text Encoding Initiative is an international project to develop guidelines for marking up electronic texts such as novels, plays, and poetry, primarily to support text analysis. As part of the mark-up a header portion has been defined, which includes metadata about the work. The TEI header, like the rest of the TEI, is defined as a Standard Generalized Mark-up Language Document Type Definition (SGML DTD).

The information in the TEI Header is similar to that captured in a library catalog. In fact, the TEI tag set can be mapped to and from MARC. In addition, elements are defined that record non-bibliographic information about the text itself, for example, how the text was transcribed or edited, what revisions have been made, and who performed the mark-up. All these metadata elements are important in text analysis and textual scholarship.

3.4 Encoded Archival Description (EAD)

Finding aids are important tools for resource description and discovery in archives and special collections of both physical and digital records. Finding aids differ from traditional library catalog records by being much longer, more narrative and explanatory, and hierarchical in their structure. The Encoded Archival Description (EAD) was developed as a way of marking up the data contained in a finding aid, so that it can be searched and displayed online.

Like the TEI Header, the EAD is defined as an SGML DTD. It begins with a header section that describes the finding aid itself (for example, who wrote it) which could be considered metadata about the metadata. It then describes the collection as a whole and successively more detailed information about the contents of the collection. Where the individual items being described exist in digital form, the EAD record can include pointers (digital identifiers) to the electronic information. This metadata scheme is particularly popular in academic libraries with large special collections and in archives. Users of EAD hope this scheme will encourage consistency and facilitate cross-archive searching. The EAD standard is maintained jointly by the Library of Congress and the Society of American Archivists.

3.5 ONIX International

ONIX (Online Information Exchange) International is a metadata scheme developed by a number of book industry trade groups in the United States and Europe to support e-commerce. ONIX has elements for basic bibliographic, trade, evaluation and promotional information for books and e-books. Although initially focused on books, ONIX has recently commissioned a study to determine the applicability of the scheme to serials.

3.6 Content Standard for Digital Geospatial Metadata

Metadata schemes for datasets are particularly significant in disciplines where numeric and statistical data are primary resources. One of the most well developed element sets and content standards for data is the U.S. Federal Geographic Data Committee's (FGDC) Content Standard for Digital Geospatial Metadata. Geospatial datasets link data for a specific purpose to the latitude and longitude coordinates on the earth. These datasets are used in a wide variety of applications, including soil and land use studies, climatology and global change monitoring, remote sensing, and demographic and social science research.

The FGDC Content Standard defines over 200 elements. The majority of these elements are optional in the standard, but they may be mandatory for specific implementations. Many national and local governments use the FGDC Content Standard; it is currently on track for incorporation as an ISO standard.

3.7 Data Documentation Initiative (DDI)

The Data Documentation Initiative is a consortium of public and private sector organizations including major universities and the U.S. Bureau of the Census. The DDI's goal is to establish metadata standards for describing social science data sets. Included are elements such as the collection method, relevant software, and units of measure. A similar initiative within the U.S. Bureau of the Census involves metadata to describe questionnaires and other survey instruments.

3.8 Technical Metadata for Still Images

The National Information Standards Organization has developed a data dictionary (July 2000, draft released for comment, February 2001) of technical elements for still images. NISO realized that the focus of most cultural institutions had been on descriptive metadata, without any emphasis on the technical aspects of digital images that would be needed to adequately store and preserve them. The purpose of the standard is to facilitate the "development of applications to validate, manage, migrate and process images of enduring value." The emphasis is not only on current use of still images, but on the long-term provenance, preservation, and assessment for use and re-use.

The Basic Image Parameters include over 50 elements. For example, there are elements that describe the format, such as compression, MIMEtype and photometric interpretation. Elements related to the image's creation include the scanning agency and camera capture settings. The change history includes the processing agency and the processing software. There are additional elements such as spatial metrics, the colormap, the image width, and the image length.

4.0 Metadata interoperability

With so many metadata schemes, how will chaos be avoided? How can we ensure that systems that use different metadata schemes will be interoperable, in other words that information collected by one organization for a particular purpose can be exchanged, transferred or used by another organization for a different purpose. Practitioners cite metadata frameworks, crosswalks, and metadata registries as ways to achieve this interoperability. However, it should be noted that there has been little large scale testing of metadata interoperability.

4.1 Metadata frameworks

A metadata framework is a reference model that provides a high-level, conceptual structure into which other metadata schemes can be placed. It also gives designers and developers a consistent, cross cutting terminology around which to discuss metadata for a particular purpose.

4.1.1. Metadata Encoding and Transmission Standard

The Metadata Encoding and Transmission Standard (METS) was developed by the Digital Library Federation and the Library of Congress for the management of digital library objects. METS uses a framework, described earlier in this paper, which defines metadata as descriptive, administrative or structural. METS also adds a fourth component, a list of the files in the digital library object. The most significant contribution of METS is its emphasis on structural metadata. The structural component of the METS scheme indicates how these files work together to form the digital library object. This not only supports the management of the object by a digital library, but it facilitates the exchange of these objects among digital libraries.

METS provides an XML DTD that can point to metadata in other schemes by declaring the scheme that is being used. For example within the METS framework, Dublin Core elements could be used to describe a digital still image for resource discovery, and the technical elements identified by NISO could be used to document the structural aspects of the image.

4.1.2 <indecs>

The Interoperability of Data in E-Commerce Systems (<indecs>) Framework is an international collaborative effort originally supported by the European Commission. It has developed a metadata framework, or a reference model, that supports the sharing of information about intellectual property rights in electronic commerce. In the basic model, people make "stuff", people use "stuff", and people make deals about "stuff."

Rather than develop a new metadata standard, <indecs> provides a framework for the various existing schemes to interact. For example, transactions related to music, journal articles or books could interchange information with one another. This framework has also been discussed as a way to allow the various groups (publishers, libraries and users) involved in access to electronic journal subscriptions to work within a consistent framework for interchange while maintaining the original metadata for their local applications.

4.1.3 Open Archive Initiative

The Open Archive Initiative (OAI) began as a project to provide consistent access across the numerous e-print services created by government and academia in the mid-1990s. However, the OAI has proven to be generally applicable for other types of electronic resources. The objective of the OAI is to create a low barrier to implementation, so OAI has only a few metadata elements based on the Dublin Core. Communities can extend the minimal set as needed.

To be OAI-compliant, the archive exposes the OAI metadata set by crosswalking its native metadata format to that of the OAI. This file is exposed and then harvested into a central repository. The software for implementing an OAI compliant archive is freely available. Communities are expected to develop their own services for searching the metadata repositories.

4.2 Metadata crosswalks

Metadata crosswalks map the elements, semantics and syntax from one metadata scheme to those of another. A crosswalk allows metadata created by one community to be used by another community that uses a different metadata standard. The degree to which crosswalks are successful depends on the similarity of the two schemes. The mapping of schemes with fewer elements (less granularity) to those with more elements (more granularity) is problematic. Despite similarity at the semantic level, the crosswalk can be problematic if the content rules differ from the original scheme to the target scheme.

However, crosswalks are important for virtual libraries and subject gateways that collect resources from a variety of sources and treat them as a whole collection, perhaps with a single search engine applied. While these crosswalks are key to interoperability, they are also labor intensive to develop and maintain.

4.3 Metadata registries

Registries are another tool for exchanging metādata. They provide information about the definition, origin, source, and location of the scheme, usage profile, element set, and/or authority files for element values. A registry maps one scheme to another so that both humans and computers can understand how they might integrate. The DESIRE (Development of a European Service for Information on Research and Education) Project funded by the European Commission has developed a prototype of such a registry based on the ISO standard for defining data elements (ISO 11179).

Registries are particularly useful in specific disciplines or industries such as health care, aeronautics, or environmental science, where they can be used to make the contents of resources more easily integrated. A good example is the U.S. Environmental Protection Agency's Environmental Data Registry which provides information about thousands of data elements used in current and legacy EPA databases. The metadata registry provides an integrating resource for legacy data, acts as a look-up tool for designers of new databases, and documents each data element.

5.0 Metadata creation

Metadata is extremely important for the discovery and management of digital resources. However, there are major issues related to the cost and time involved in creating this metadata. A variety of methods are used for creating metadata from manual creation to metadata creators/editors and metadata generators.

5.1 Manually created metadata

Who creates metadata? The answer to this varies by discipline, the electronic information being described, the tools available, and the expected outcome. In the case of descriptive metadata, originators may provide some level of metadata creation. This is particularly true in the documentation of datasets where the originator has significant understanding of the rationale for the dataset and the uses to which it could be put, and where there is little if any textual information that a cataloger could use. In other cases, projects have found that it is necessary to have metadata catalogers or librarians create the descriptive metadata or at least review the metadata created by the originators, because the originators do not have the time or the skills to create adequate metadata.

The resources involved in creating metadata for the wealth of electronic information has led to the development of two types of tools -- metadata generators and metadata creators/editors. Creators/editors support the manual creation and editing of metadata. Metadata generators automatically create a metadata record based on the HTML from a Web site.

5.2 Metadata creators/editors

Metadata creators/editors, both commercial and proprietary, address the need for speed and quality. Many tools support validation rules and pick lists based on authority files or controlled domains, including controlled vocabularies and thesauri. Templates may be provided and customized to stream line the data entry process.

There are several metadata creators/editors for the Dublin Core. The Nordic Web provides metadata creation software and Dublin Core to MARC conversion software, which is free within the European Union. MetaWeb from Australia has developed a metadata editor called "Reggie."

There are a number of FGDC- compliant metadata creation tools, including Metamaker, which was developed by the U.S. Geological Survey, Biological Resource Division. Some FGDC-compliant products have been developed by geographic information system (GIS) vendors to support the documentation of created or stored within their products. While many of these systems are proprietary, there are efforts underway through the Open GIS Consortium to support an open metadata tool.

5.3 Metadata generators

The program DC.doc from the UK Online Library Network (UKOLN) analyzes a Web site (indicated by a URL that the user provides) and creates a Dublin Core record. The proposed content is displayed back to the user in a Dublin Core template. The user can modify the content. The DC fields can be returned to the Web site as metatags or stored in a separate file. Similar programs are available from The Nordic Web Project, OCLC's Cooperative Online Resource Catalog system for shared cataloging of Web resources, and Australia's MetaWeb.

Of course, the content of metadata generators is only as good as the content of the originating Web site. None of these tools provide 100% automatic metadata generation, particularly if high quality content is desired. Users often use the software simply for a handy Dublin Core template.

6.0 Controlled vocabularies and metadata

The use of controlled vocabularies is becoming increasingly important as a tool for metadata creation and access. This is particularly true as more information managers realize the problems that arise from free text searching or the use of uncontrolled keywords.

Most metadata schemes do not dictate the use of a particular controlled vocabulary when entering the contents of elements that describe what the resource is about. However, use of controlled vocabularies is encouraged, particularly within a subject domain. Many metadata schemes allow controlled vocabularies to be defined within the syntax.

A variety of controlled vocabulary systems are being used for indexing electronic resources. These include traditional library schemes such as the Library of Congress Subject Headings and the Dewey Decimal Classification, specific domain-oriented thesauri or classification schemes, and locally created lists of frequently used or important terms. The tools that are required to use existing controlled vocabulary schemes in the Internet environment is a major research area for OCLC.

Individual projects may specify the controlled vocabularies to be used. For example, the National Biological Information Infrastructure, which uses the biological profile for the FGDC Geospatial Content Standard, specifies the controlled vocabulary to be used. Cambridge Scientific Abstracts, as a partner of the National Biological Information Infrastructure (NBII), is developing a Biocomplexity Thesaurus. The terms in the thesaurus will be used in the NBII metadata to tag electronic resources across the NBII subject and geographic nodes. The thesaurus will also be used to select terms for the more traditional bibliographic indexing in CSA's Biocomplexity database, which is searchable through the NBII Web site. The NBII portal will use the terms to create collections of information based on a user's personal preferences. The NBII's biological profile of the FGDC Metadata Content Standard also specifies the use of the Integrated Taxonomic Information System as the authority file for completing the biological taxonomic classification elements within the metadata record.

Controlled term lists have been developed by many of the U.S. states using the Global Information Locator Service. These include terms that describe the major services and products provided by states to their citizens, to state employees, or to other governments, whether state, local or national.

Unfortunately, the use of a variety of controlled vocabulary schemes does not significantly improve searching across the breadth of Internet resources or when the user is searching outside his or her area of expertise. A group called the Networked Knowledge Organization Systems/Services (NKOS), an ad hoc group from public and private sector organizations in ten countries, has been discussing the issues related to providing generally applicable knowledge organization services (KOS) via the Internet. The group defines KOSs to include authority files, thesauri, gazetteers, ontologies, topic maps, taxonomies, subject headings, and any other type of scheme intended to organize digital objects. NKOS has been developing protocols for the use of KOSs via the Internet, and has developed a set of metadata elements to describe KOSs and their behavior. This metadata could be used as part of a registry of KOSs or as metatag information embedded in header information for a Web-based KOS. NKOS and the U.S. National Information Standards Organization (NISO) are discussing how to advance work on these projects, specifically for electronic thesauri.

In a similar initiative, a Z39.50 profile for thesauri has been developed. The profile provides a high level, abstract representation for navigating a thesaurus. In addition to providing thesaurus search capabilities within the realm of Z39.50 (which includes the GILS initiatives and many of the initiatives that use the FGDC content standard) an appendix to the profile provides an XML DTD for thesauri that could be used by other protocols.

7.0 Conclusions

All metadata share a common purpose – resource discovery, location and collection organization. However, because the needs of resource types and user communities differ, many schemes have been developed, along with specific extensions and profiles. Metadata standards and interoperability remain key issues. In order to increase the use of metadata, systems need to be developed that support metadata creation at the same time that the resource is created. Larger testbeds of metadata and search engines that take more advantage of metadata that has been created must be developed. Communities of practice need to develop content standards and to look for common areas of interest in order to support access to information across communities. Most importantly, creators of electronic resources must be made aware of the importance of metadata for the short as well as the long-term use of their contributions to the world of electronic information resources.

8.0 Selected resources on metadata, frameworks and related standards[1]

General Resources about Metadata

Distributed Systems Technology Centre. (2000). Metadata Schema Registry (Australia). Retrieved May 3, 2002 from the Metadata Schema Registry Web site: metadata.net/

Hodge, G. (2001). Metadata Made Simpler. Retrieved May 3, 2002 from the National Information Standards Organization Web site: www.niso.org/news/Metadata_simpler.pdf

International Federation of Library Associations and Institutions (IFLA). (2002). Digital Libraries: Metadata Resources. Retrieved May 3, 2002 from the International Federation of Library Associations and Institutions Web site: www.ifla.org/II/metadata.htm

Metadata Information Clearinghouse (Interactive). (1999). Retrieved May 3, 2002 from the Metadata Information Clearinghouse Web site: www.metadatainformation.org/

Schwartz, C. (2002). Metadata Portals & Multi-standard Projects. Retrieved May 3, 2002 from the Simmons College Web site: web.simmons.edu/~schwartz/meta.html

UK Online Library Network (UKOLN). (2002). Metadata Resources. Retrieved May 3, 2002 from the UK Online Library Network Web site: www.ukoln.ac.uk/metadata/resources

Selected Metadata Schemes and Frameworks

Data Documentation Initiative: DDI. Retrieved May 3, 2002 from the University of Michigan Web site: www.icpsr.umich.edu/DDI/

DESIRE (Development of a European Services for Information on Research and Education). (2002). Metadata Registry. Retrieved May 3, 2002 from the UK Online Library Network Web site: desire.ukoln.ac.uk/registry/

Dublin Core Metadata Initiative. (2002). Retrieved May 3, 2002 from the OCLC Web site: purl.oclc.org/metadata/dublin_core/

Ellerman, Castedo. (1997). Channel Definition Format (CDF). Retrieved May 3, 2002 from the WWW Consortium Web site: www.w3.org/TR/NOTE-CDFsubmit.html

Encoded Archival Description (EAD). (2001). Retrieved May 3, 2002 from the Library of Congress Web site: lcweb.loc.gov/ead/

FGDC Content Standard for Digital Geospatial Metadata (CSDGM). (2001). Retrieved May 3, 2002 from the Federal Geographic Data Committee Web site: www.fgdc.gov/metadata/contstan.html

GEM (Gateway to Educational Materials) Element Set & Profile(s) Workbench. (2002). Retrieved May 3, 2002 from the GEM Web site: www.geminfo.org/Workbench/Metadata/index.html

Global Information Locator Service (GILS). Retrieved May 3, 2002 from the Global Information Locator Service Web site: www.gils.net

[1] Inclusion in this list does not constitute endorsement by Information International Associates or the U.S. Geological Survey.

IMS Global Learning Consortium, Inc. (2002). Learning Resource Meta-data Specification. Retrieved May 3, 2002 from the IMS Global Learning Consortium Web site: www.imsproject.org/metadata/index.html

<indecs> interoperability of data in e-commerce systems. Retrieved May 3, 2002 from the <indecs> Web site: www.indecs.org/

METS: Metadata Encoding and Transmission Standard. (2001). Retrieved May 3, 2002 from the Library of Congress Web site: www.loc.gov/standards/mets/

OCLC/RLG working Group on Preservation Metadata. (2001). A Recommendation for Content Information. Retrieved May 3, 2002 from the OCLC Web site: www.oclc.org/research/pmwg/contentinformation.pdf

ONIX (Online Information Exchange). Retrieved May 3, 2002 from the Editeur Web site: www.editeur.org/

Technical Metadata for Digital Still Images. (2001). Retrieved May 3, 2002 from the National Information Standards Organization Web site: www.niso.org/committees/committee.au.html

TEI Consortium. Text Encoding Initiative. (2002). Retrieved May.3, 2002 from the TEI Web site: www.tei-c.org/

Visual Resources Association Data Standards Committee. VRA Core Categories, Version 3.0. (2002). Retrieved May 3, 2002 from the VRA Web site: www.vraweb.org/vracore3.htm

Metadata Crosswalks

Day, M. (2001). Metadata: Mapping between Metadata Formats
(comprehensive list of mappings to and from all major formats including national versions of MARC) Retrieved May 3, 2002 from the UK Online Library Network Web site: www.ukoln.ac.uk/metadata/interoperability/

St. Pierre, M. and W. La Plant. Issues in Crosswalking Content Metadata Standards. (1998). Retrieved May 3, 2002 from the National Information Standards Organization Web site: Niso.org/press/whitepaper/crsswalk.html

Metadata Tools

BlueAngel Technologies (MetaStar). Retrieved May 3, 2002 from the BlueAngel Technologies Web site: www.blueangeltech.com/

Distributed Systems Technology Centre. (1999). MetaWeb Project (Australia). Retrieved May 3, 2002 from the Distributed Systems Technology Centre Web site: www.dstc.edu.au/RDU/MetaWeb

Dublin Core Metadata Initiative. (2002). Tools and Software. Retrieved from the Dublin Core Web site: dublincore.org/tools/

Federal Geographic Data Committee. Metadata Tools. Retrieved May 3, 2002 from the Federal Geographic Data Committee Web site: www.fgdc.gov/metadata/metatool.html

Intergraph Spatial Metadata Management System. (2002). Retrieved May 3, 2002 from the Intergraph Web site: www.intergraph.com/gis/smms//

Meta Matters. Retrieved May 3, 2002 from the National Library of Australia Web site: www.nla.gov.au/meta/

Metadata: UKOLN Software Tools.
(comprehensive list of tools for a variety of standards including Dublin Core, GILS and IMS)
Retrieved May 3, 2002 from the UK Online Library Network Web site: www.ukoln.ac.uk/metadata/software-tools/

MetaPackager. (2002). Retrieved May 3, 2002 from the HiSoftware Web site: www.hisoftware.com/metapackager.htm

Nordic Metadata Projects. (2000). Retrieved May 3, 2002 from the University of Helsinki Library Web site: www.lib.helsinki.fi/meta/

Related Initiatives

Australian Government Locator Service. (2000). Retrieved May 3, 2002 from the Australia Government Online Web site: www.govonline.gov.au/projects/standards/agls.html

CORC (Cooperative Online Resource Catalog). (2002). Retrieved May 3, 2002 from the OCLC Web site: www.oclc.org/corc/about/

CrossRef. (2000). Retrieved May 3, 2002 from the CrossRef Web site: www.crossref.org

National Biological Information Infrastructure (U.S.). (2001). Retrieved May 3, 2002 from the NBII Web site: www.nbii.gov

National Spatial Data Infrastructure (U.S.) (2002). Retrieved May 3, 2002 from the Federal Geographic Data Committee Web site: www.fgdc.gov/nsdi/nsdi.html

Networked Knowledge Organization Systems/Services (NKOS). (2002). Retrieved May 3, 2002 from the School of Library and Information Science, Kent State University Web site: nkos.slis.kent.edu/

Open Archives Initiative. (2002). Retrieved May 3, 2002 from the OAI Web site: www.openarchives.org/

Related Standards

Corporation for National Research Initiatives. Handle System®. (2002). Retrieved May 3, 2002 from the Handle Web site: www.handle.net/

Extensible Markup Language (XML). (2002). Retrieved May 3, 2002 from the WWW Consortium Web site: www.w3.org/XML/

International DOI Foundation. DOI: Digital Object Identifier System. (2002). Retrieved May 3, 2002 from the International DOI Foundation Web Site: www.doi.org/

PURL (Persistent URL). (2002). Retrieved May 3, 2002 from the PURL Web site: purl.org

Resource Description Framework. (2002). Retrieved May 3, 2002 from the WWW Consortium Web site: www.w3.org/RDF/

Taylor, M. Zthcs: a Z39.50 Profilc for Thcsaurus Navigation, Vcrsion 0.4. (2000). Rctrievcd May 3, 2002 from the Library of Congress Web site: lcweb.loc.gov/z3950/agency/profiles/zthes-04.html

W3C: World Wide Web Consortium. (2002). Retrieved May 3, 2002 from the WWW Consortium Web site: www.w3.org/

Z39.50. (2002). Retrieved May 3, 2002 from the Library of Congress Web site: ww.loc.gov/z3950/agency/